LARNet: Lie Algebra Residual Network for Face Recognition

Xiaolong Yang 1 2 3  Xiaohong Jia 1 2  Dihong Gong 3  Dong-Ming Yan 4 2  Zhifeng Li 3  Wei Liu 3

Abstract

Face recognition is an important yet challenging problem in computer vision. A major challenge in practical face recognition applications lies in significant variations between profile and frontal faces. Traditional techniques address this challenge either by synthesizing frontal faces or by pose invariant learning. In this paper, we propose a novel method with Lie algebra theory to explore how face rotation in the 3D space affects the deep feature generation process of convolutional neural networks (CNNs). We prove that face rotation in the image space is equivalent to an additive residual component in the feature space of CNNs, which is determined solely by the rotation. Based on this theoretical finding, we further design a Lie Algebraic Residual Network (LARNet) for tackling pose robust face recognition. Our LARNet consists of a residual subnet for decoding rotation information from input face images, and a gating subnet to learn rotation magnitude for controlling the strength of the residual component contributing to the feature learning process. Comprehensive experimental evaluations on both frontal-profile face datasets and general face recognition datasets convincingly demonstrate that our method consistently outperforms the state-of-the-art ones.

1. Introduction

The recent development of deep learning models and an increasing variety of datasets have greatly advanced face recognition technologies (Liu et al., 2017; Wang et al., 2018a; Deng et al., 2019). Although many deep learning models are strong and robust to face recognition conducted in unconstrained environments, there remain quite a lot of challenges for recognizing faces across different age levels (Gong et al., 2013a; Wang et al., 2019; 2018b; Gong et al., 2015; Li et al., 2016b), different modalities (Li et al., 2014; Gong et al., 2017; Li et al., 2016a; Gong et al., 2013b; Luo et al., 2021), different poses (Huang et al., 2000; Cao et al., 2018a; Masi et al., 2016a; AbdAlmageed et al., 2016), and occlusions (Song et al., 2019; Zhang et al., 2007). In this paper, we develop a robust recognition algorithm to address the challenges for general face recognition, with a particular effect on matching faces across different poses (e.g., frontal vs. profile). Since the generalization ability of the deep model is closely related to the size of the training data, given an uneven and insufficient distribution of frontal and profile face images, the deep features tend to focus on frontal faces, and the learning results are only biased incomplete statistics. In order to tackle this problem, some work has reconstructed more datasets by different data augmentation methods. A typical way is to enrich input sources either by the synthesis of profile faces with appearance variations (Masi et al., 2016b) or by a set of images as one image input (Xie & Zisserma, 2018), so that the need for profile data is alleviated. Another way is to combine more data information, including multi-task learning (AbdAlmageed et al., 2016; Masi et al., 2016a; Yin & Liu, 2017) and template adaptation (Hassner et al., 2016; Crosswhite et al., 2018). Here, multi-task learning focuses on pose-aware targets, combined with richer information such as illumination, expression, gender, and age, to comprehensively boost the recognition performance; while the method based on template adaptation learning always creates a mean 3D model face, and by means of migration and mapping it avoids processing the 3D transformation at the image level. Nevertheless, these strategies tend to increase the unnecessary computational burden. Some other approaches use profile faces to synthesize frontal faces so that they can avoid large pose variations (Tran et al., 2017; Wang et al., 2017; Yin et al., 2017; Zhou et al., 2020). However, these methods would suffer from artifacts caused by occlusions and non-rigid expressions.

The above mentioned work mostly relies on additional data sources or additional labels. A recent new work called Deep Residual Equivalent Mapping (DREAM) (Cao et al., 2018a)
We prove that for each frontal-profile pair linked by a rotation, and we prove that a face rotation in the image space is equivalent to an additive residual component in the deep feature space. To show the equivalence, we reconstruct the image corresponding to the modified feature (blue dot) and provide the visual result for the expected frontal face.

We observe a natural fact that frontal-profile pairs are generated by head rotation, which should not be ignored in profile face recognition. However, rotation matrices are not easy to be embedded in CNNs. This is because the group of rotation matrices is closed under multiplication but not closed under addition, while the addition operation appears frequently in all gradient descent calculations. Benefiting from the pose estimation work (Tuzel & abd Peter Meer, 2008) in the field of simultaneous localization and mapping (SLAM), we develop a novel approach of using Lie algebra to achieve the updates of rotation matrices in CNNs.

We prove that for each frontal-profile pair linked by a rotation, their corresponding deep features also preserve a corresponding rotation relationship. To the best of our knowledge, this is the first attempt to theoretically explore and explain the physical relationship between the features of a frontal face and its profile counterpart. Based on this theoretical result, we propose the Lie Algebra Residual Network (LARNet). LARNet achieves face frontalization or rotation-and-render in the feature space, as shown in Fig. 1. Meanwhile, we conduct comparative experiments with more than 30 solutions under various evaluation criteria and metrics, while our method outperforms representative state-of-the-art competitors. In summary, our contributions are three-fold:

1. We theoretically prove that the features of frontal-profile pairs have a physical relationship based on rotations in a residual network using Lie algebra, which is equivalent to an additive residual component in the deep feature space of CNNs.

2. We design a novel gating subnet based on the proof, which neither needs to modify the original backbone network structure nor relies on a large number of modules, but brings a great performance improvement.

3. LARNet enhances the deep model’s ability in feature representation and feature classification, and accomplishes superior performance on various datasets and in various evaluation criteria, including frontal-profile face verification-identification and general face recognition tasks.

2. Related Work

We briefly discuss the most related work of profile face recognition and large-pose face recognition.

**Insufficient Dataset.** Many methods try to solve the profile face recognition problem by avoiding the unevenness of datasets. Masi et al. (2016b) proposed domain-specific data augmentation with increasing training data sizes for face recognition systems, and focused on important facial appearance variations. Multicolour Network (Xie & Zisserma, 2018) and Neural Aggregation Network (NAN) (Yang et al., 2017) propose to use more information, such as a set of images or videos as input, to tackle the potential shortcomings of merely using a single image. These methods are not easy to avoid falsely matching profile faces of different identities or missing frontal and profile faces of the same identity.

**Pose Variation.** Many existing methods have conducted in-depth researches on large poses. Template-adaptation-based work (Hassner et al., 2016; Crosswhite et al., 2018) is mainly engaged with transfer learning by means of a constructed classifier and synthesizer, and pooling based on image quality and head poses. As opposed to those techniques which expect a template model to learn pose invariance, pose aware deep learning methods (AbdAlmageed et al., 2016; Masi et al., 2016a) use multiple pose-specific models and rendered face images, which reduce the sensitivity to pose variations. More work favors using more labels instead of pose itself. Multi-task learning (MTL) has been widely used, which consists of pose, illumination, and expression estimations. Yin et al. (2017) proposed a pose-directed multi-task CNN and united the balance between different tasks. DebFace (Zhou et al., 2020) (de-biasing adversarial network) additionally takes gender, age, and race into consideration, and minimizes the correlation among feature factors so as to abate the bias influence from the other factors. These methods are effective; however, the use of multiple models and tasks tends to increase the computational cost, and the accuracy of their results is confined.
**Frontalization.** Since profile and large-pose faces bring more challenges, some methods directly use an available dataset to synthesize frontal faces to perform face recognition. Due to the widespread use of GAN, FF-GAN (Yin & Liu, 2017) and DR-GAN (Tran et al., 2017) surpass the performances of many competitors, with disentangled encoder-decoder structure towards learning a generative and discriminative representation. With the rapid progress of 3D face reconstruction technology, projecting rendering of frontal faces after reconstruction has also risen. Rotate-and-Render (Zhou et al., 2020) is a representative work from single-view images, and can leverage the recent advances in 3D face modeling and high-resolution GAN to constitute building blocks, since the 3D rotation-and-render of faces can be applied to arbitrary angles without losing facial details. Note that the reconstruction and synthesis approaches have their advantages in visualization performance, but their feature representation capability is inadequate for applications in unconstrained environments.

**Feature Space.** Some work considered features rather than image itself. Shi et al. (2019) proposed Probabilistic Face Embeddings (PFEs), which represent each face image as a Gaussian distribution in the latent space. Feature Transfer Learning (Yin et al., 2019) encourages the under-represented distribution close to a Gaussian prior. Although both DREAM and their ad-hoc designed results reach a bottleneck of feature-representation capability is inadequate for applications in unconstrained environments.

3. Methodology

In this section, we assume that a frontal face and its profile face have a corresponding rotation relationship in the original 3D space. For ease of understanding, only the rotation with the orthogonal transformation relationship is discussed here. The derivation of more complex Euclidean transformation relationships, including translation and zooming, is referred to our supplementary material.

3.1. Problem Formulation

Our goal is to find a transformation between the features of an input profile face image and the expected frontal face image, to realize frontalization in the deep feature space and to achieve a powerful feature representation robust to pose variations, as shown in Fig. 1.

We denote $\mathcal{F}(x)$ as a feature extraction function in CNNs for an input image $x$. Here, for each pixel $(u, v)$ in the image $x$, we adopt its homogeneous coordinate representation $(u, v, 1)^T$, and for convenience still denote the collection of these 3D homogeneous coordinates as $x$.

Let $d$ be the dimension of layers to be considered. Then the extracted feature $\mathcal{F}(x) \in \mathbb{R}^d$. We shall prove that there exists a map $\mathcal{R}_{map}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that acts as a rotation in the deep feature space corresponding to the rotation $R \in SO(3)$ of the (homogenized) image $x$:

$$\mathcal{F}(R \cdot x) = \mathcal{R}_{map}(\mathcal{F}(x)).$$

For the frontal face image $x_f$ and its profile face image $x_p$, the homography transformation matrix of these two images degenerates into a rotation matrix: $x_f = R \cdot x_p$, and we have:

$$\mathcal{F}(x_f) = \mathcal{F}(R \cdot x_p) = \mathcal{R}_{map}(\mathcal{F}(x_p)).$$

Furthermore, we try to use Lie group theory (Rossmann, 2002) and prove that the mapping $\mathcal{R}_{map}(\cdot)$ can be decomposed into an additive residual component, which is solely determined by the rotation:

$$\mathcal{F}(x_f) = \mathcal{F}(x_p) + \omega(R) \cdot \mathbf{C}_{res}(R, x_p).$$

Thus, we only need a residual subnet $\mathbf{C}_{res}$ for decoding pose variant information from the input face image, and a gating subnet $\omega$ to learn rotation magnitude for controlling the strength of the residual component contributing to the feature learning process. Eq. (3) is the core principle of LARNet we propose, and the detailed derivations and experimental design will be presented in the following sections.

3.2. Rotation in Networks and Lie Algebra

To find what $\mathcal{R}_{map}$ exactly is, we have tried to directly explore and analyze the role of rotation $R$ in networks from Eq. (2). From the original paper of ResNet (He et al., 2016), it proposes a novel shortcut, which not only retains the
depth of deep networks, but also has the advantages of shallow networks in avoiding the overfitting issue. The feature learning from shallow layer $l$ to deep layer $L$ is described as:

$$x_L = x_l + \sum_{i=1}^{L-1} H(x_i, w_i),$$

(4)

Next, we show the addition and multiplication in Lie algebra by Baker-Campbell-Hausdorff (BCH) formula (Wulf, 2002; Brian, 2015) and Friederichs’ theorem (Wilhelm, 1954; Jacobson, 1966):

$$\exp(\Delta\phi^\wedge) \exp(\phi^\wedge) = \exp\left((\phi + J_l(\phi)^{-1}\Delta\phi^\wedge)\right),$$

$$\exp((\phi + \Delta\phi^\wedge)) = \exp\left([J_l\Delta\phi^\wedge]\exp(\phi^\wedge)\right).$$

(10)

$$J_l$$ is the left Jacobian of $SO(3)$. For a point $p \in \mathbb{R}^3$, the derivative of $Rp$ with respect to a perturbed rotation is:

$$\frac{\partial (Rp)}{\partial (\Delta \phi)} = \lim_{\Delta \phi \to 0} \frac{\exp(\Delta \phi^\wedge) \exp(\phi^\wedge) p - \exp(\phi^\wedge) p}{\Delta \phi} = -(Rp)^\wedge.$$  

(11)

For a current $R_i$, we choose a perturbation $\Delta \phi^\wedge$, such that $R_{i+1} = \exp(\Delta \phi^\wedge)R_i$. Then for a point $p$, Eq. (11) leads:

$$R_{i+1}p = \exp(\Delta \phi^\wedge)R_ip \approx R_ip - (R_ip)^\wedge \Delta \phi.$$  

(12)

Then for the target function that is to be optimized, denoted by $u$, we use Taylor expansion to write:

$$u(R_{i+1}p) = u(\exp(\Delta \phi^\wedge)R_ip) \approx u((1 + \Delta \phi^\wedge)R_ip) \approx u(R_ip) - \frac{\partial u}{\partial d} \bigg|_{d=R_ip} (R_ip)^\wedge \Delta \phi = u(R_ip) + \delta^T \Delta \phi.$$  

(13)

We need to determine $\Delta \phi$ such that the value of $u$ decreases. A possible choice is to pick $\Delta \phi = -\alpha D\delta$, where $\alpha > 0$ is a small step size and $D$ is an arbitrary positive-definite matrix. Applying this perturbation within the scheme, we can update the rotation matrix by $R_{i+1} \leftarrow \exp(-\alpha D\delta^\wedge)R_i$.

Back to the original problem, given Eqs. (10-13), we can rewrite the first principal term of Eq. (5) as follows:

$$\frac{\partial \text{Loss}}{\partial x_L^f} \approx \frac{\partial \text{Loss}}{\partial x_L} \frac{\partial x_L}{\partial x_L} = \frac{\partial \text{Loss}}{\partial (R \cdot x_L^f)} \cdot \partial \Delta \phi$$

$$= -(R \cdot x_L^f)^\wedge \cdot \partial \Delta \phi$$

(14)

Note that in Eq. (2), we mentioned that the homography relationship between the two original images $x_p$ and $x_f$ is connected by a rotation, but this relationship generally cannot be guaranteed in the CNNs. However, Eq. (14) suggests
that this relationship is inherited in another way during gradient decent at each layer. In fact, since $\mathbf{R} \in SO(3)$, $\mathbf{R} \cdot x_p^L$ and $x_p^f$ are asymptotically stable according to Lyapunov’s second method (Lukyanov, 1992; Bhatia & Szegö, 2002). With the gradual training progress of the ResNet, the feature vectors of $\mathbf{R} \cdot x_p^L$ and $x_p^f$ have the same convergent representation: $\mathcal{F}(x_f) = \mathcal{F}(\mathbf{R} \cdot x_p)$. Furthermore, we decouple the rotation relation from face features into Eq. (9) and Eq. (12).

Let $V_{\text{res}} = \mathcal{F}(\mathbf{R} \cdot x_p) - \mathcal{R}_{\text{map}} \mathcal{F}(x_p) \in \mathbb{R}^d$ be the residual vector, and we have:

$$
\mathcal{R}_{\text{map}}^{-1}(x_f) = \mathcal{F}(x_p) + \mathcal{R}_{\text{map}}^{-1}(V_{\text{res}}),
$$

$$
\mathcal{F}(x_f) = \mathcal{F}(x_p) + \mathcal{R}_{\text{map}}^{-1}(V_{\text{res}} + \mathcal{R}_{\text{map}} \mathcal{F}(x_f) - \mathcal{F}(x_f)).
$$

(15)

Since during our training stage, the feature $\mathcal{F}(x_p)$ is approaching to $\mathcal{R}_{\text{map}} \mathcal{F}(x_f)$ (we shall show a corresponding analysis in Eq. (17) in Sec. 3.3), Eq. (15) leads to Eq. (16):

$$
\mathcal{F}(x_f) \approx \mathcal{F}(x_p) + \mathcal{R}_{\text{map}}^{-1}(\mathcal{F}(x_p) - \mathcal{R}_{\text{map}} \mathcal{F}(x_p)).
$$

(16)

This agrees exactly with Eq. (3). Hence, we can design the gating control function $\omega(\mathbf{R})$ as $\mathcal{R}_{\text{map}}^{-1}$ to filter the feature flow, and the component $C_{\text{res}}(\mathbf{R}, x_p) = \mathcal{F}(x_p) - \mathcal{R}_{\text{map}} \mathcal{F}(x_p)$ is obtained through residual network training.

### 3.3. The Architecture of Subnet

As previously stated, we expect to design a residual subnet $C_{\text{res}}$ for decoding pose variant information from input face images. The residual formulation in Eq. (3) allows us to use a succinct enough network structure for learning the residual compensation from the clean deep features, which is a relatively easy task. The most convenient way is to add the gating control residuals directly to the existing backbone (Arcface (Deng et al., 2019) with ResNet-50 in our paper).

Residual learning can be arranged before the final fully-connected (FC) layer of the ResNet-50 without revising any learned parameters of the original backbone model. Our residual learning has two fully-connected layers with Parametric Rectified Linear Unit (PReLU) (He et al., 2015) as the activation function. We train it by minimizing $\ell_2$ norm of the difference between the profile features $\mathcal{F}(x_p)$ and frontal features under the rotation $\mathcal{R}_{\text{map}} \mathcal{F}(x_f)$ using stochastic gradient descent.

$$
\min_{\Omega_p} \sum \left\| \mathcal{F}(x_p) - \mathcal{R}_{\text{map}}(\Omega_p) \mathcal{F}(x_f) \right\|_2^2,
$$

(17)

where $\Omega_p$ denotes the learnable parameters. We train this subnet on frontal-profile pairs sampled from the MS-Celeb-1M dataset (mentioned in Sec. 4.2), and fix these parameters for the testing. Applying a subnet with complicated structure may increase the risk of overfitting, and the design with two FC layers is on the consideration of both the task difficulty and the risk of model robustness.

Furthermore, we design a gating control function $\omega$ to analyze the rotation magnitude for controlling the strength of the residual component contributing to the deep feature learning process. In our problem context, $\omega$ needs to satisfy the following properties:

- $\omega \in [0, 1]$. Intuitively, when the input is frontal face $x_0$, there is almost no difference in the feature representation in the same network, and $C_{\text{res}}$ of residual learning will bring errors and weaken the classification ability. Therefore, it is expected that the gating control function is 0 in this case; ideally, the magnitude of the residual is thus the largest at the complete profile pose: $\mathcal{F}(x_0) - \mathcal{F}(x_{\pi/2})$, so the maximum value of the gating control function is 1:

$$
\mathcal{F}(x_0) = \mathcal{F}(x_{\pi/2}) + 1 \ast (\mathcal{F}(x_0) - \mathcal{F}(x_{\pi/2})) = \mathcal{F}(x_0).
$$

- $\omega$ has symmetric weights. A gating control function learns rotation magnitude for controlling the strength of the residual component contributing to the feature learning process, and the same deflection angles should bring the same influence on frontal faces (for example, yaw angle with turning left or right). We will also use data flipping augmentation to strengthen the symmetry of the model during training.

Besides, it is worth noting that roll, yaw, and pitch angles have different contributions to the final face recognition performance. The effect of roll will be eliminated by face alignment (mentioned in Sec. 4.2), while face images with
false positive pair
frontal face data
profile face data

Figure 3. The effect of our gating control function for different identities. This is a challenging false positive example for a general face recognition model (Wang et al., 2018a). We collect more frontal and profile face data of those two individuals, and visualize feature vectors of all the images corresponding to this sample. It is obvious that our subnet with the gating control function is instrumental in obtaining better classification and clustering performance.

4. Experimental Results

In this section, we first provide a description of implementation details (Sec. 4.1). Besides, we list all the datasets used in the experiments and briefly explain their own characteristics (Sec. 4.2). Furthermore, we present two ablation studies on the architecture and gating control function, respectively, which explain the effectiveness of our experimental design on recognition performance (Sec. 4.3). We also compare with existing methods and some findings about profile face representation, and conduct extensive experiments on frontal-profile face verification-identification and general face recognition tasks (Sec. 4.4).

4.1. Implementation Details

Network Architecture. (1) Backbone: our backbone network is ResNet with a combined margin loss: CM(1, 0.3, 0.2) proposed by ArcFace. According to recent researches, i.e., Saxe et al. (2014), Highway Networks (Srivastava et al., 2015), and Balduzzi et al. (2017), ResNet-50 has the best layers balancing efficiency and accuracy. Therefore, we choose the ResNet with 50 layers as it is very popular in a large amount of existing work and also convenient to compare. (2) LARNet: after the backbone, the clean deep features are mapped to rotation through the subset and gating control function. (3) LARNet+: similar to DREAM (Cao et al., 2018a), we also use the end-to-end mechanism to further improve the performance of our results. Based on LARNet, we make residual learning together with the backbone network in an end-to-end manner. We train the ResNet and residual learning together and then train the residual learning separately with pose-variant frontal-profile face pairs.

Data Preprocessing. As shown in Fig. 4, we use MTCNN (Zhang et al., 2016) to detect face areas and facial landmarks on both training and testing sets. We use a flipping strategy to achieve data augmentation and enhance our model’s ability to learn symmetry. In addition, face alignment and scaling (224 x 224) are taken into account to reduce the impact caused by translation and zooming, when we only consider SO(3) instead of SE(3).

Training Details. The model is trained with 180K iterations. The initial learning rate is set as 0.1, and is divided by...
Figure 4. Data prepossessing on frontal and profile faces. (a) Face flipping: data enhancement for strengthening the model’s ability to learn symmetry; (b) face alignment: reduce the impact caused by translation and rotation in a plane such that the eyes lie along a horizontal line; (c) face scaling: reduce the impact caused by zooming because focal length differs in every image such that faces are approximately identical in size.

10 after 100K, 160K iterations. The SGD optimizer has momentum 0.9, and weight decay $5e_{−4}$. For rotation angles, we take rotation estimation via state-of-the-art work (Yang et al., 2019; 2020) and obtain pose labels. The protocols of training and testing for different datasets will be explained in the next section.

### 4.2. Datasets Exhibition

**Training Data.** We separately employ the two most widely used face datasets as training data in order to conduct fair comparison with the other methods, i.e., cleaned MS-Celeb-1M database (MS1MV2) (Guo et al., 2016) and CASIA-WebFace (Yi et al., 2014). MS1MV2 is a clean version of the original MS-Celeb-1M face dataset that has too many mislabeled images, containing 5.8M images of 85,742 celebrities. CASIA-WebFace uses tag-similarity clustering to remove noise of the data source, containing 500K images of 100K celebrities from IMDb.

**Testing Data.** We explore several efficient face verification datasets for testing. Celebrities in Frontal-Profile (CFP) (Sengupta et al., 2016) is a challenging frontal to profile face verification dataset, containing 500 celebrities, each of which has 10 frontal and 4 profile face images. We further test another challenging dataset IARPA Janus Benchmark A (IJB-A) (Klare et al., 2015) that covers extreme poses and illuminations, containing 500 identities with 5,712 images and 20,414 frames extracted from videos. Besides focusing on frontal-profile face verification, we also conduct experiments on general face recognition datasets to verify that our method can reach the state of the art on general face recognition tasks. Including the most widely used LFW (Huang et al., 2008) dataset (13,233 face images from 5,749 identities) and YTF (Wolf et al., 2011) dataset (3,425 videos of 1,595 different people), we also report the performance of Cross-Pose LFW (CPLFW) (Zheng & Deng, 2018), which deliberately searches and selects 3,000 positive face pairs with pose differences to add pose variations to intra-class variance, so the effectiveness of several face verification methods can be fully justified. Furthermore, we also extensively conduct a more in-depth ablation experiment on the Large-scale CelebFaces Attributes (CelebA) dataset (Liu et al., 2015b), containing 10,177 celebrities and 202,599 face images, which covers large pose variations.

### 4.3. Ablation Studies

To justify that our LARNet does improve the performance of profile face recognition, we conduct two ablation experiments: 1) the architectures with/without the gating control function, and 2) the forms of the gating control function.

#### 4.3.1. The Architecture of LARNET

We study the effectiveness of architectures with and without the gating control function as well as an end-to-end optimization manner. All the results are using the same backbone ResNet-50, combined margin loss in Arcface (Deng et al., 2019), and MS1MV2 training dataset, and evaluation is conducted on the CFP-FP dataset. The baseline method is Arcface without refinement. As shown in Table 1, we observe that compared with the strong baseline without the gating control function, our LARNet brings a significant advancement by 5.88% in verification accuracy. LARNet+ with the end-to-end optimization manner also plays an important role in further performance improvement, and achieves a more superior result 99.21% than LARNet 98.84%.

#### 4.3.2. The Gating Control Function

Next, we further analyze which type of gating control function has a greater contribution to the performance. For facilitating a fair comparison, all methods take the same CASIA-WebFace dataset and ResNet-50 backbone for training, and evaluation is conducted on CelebA with metric Equal Error Rate (EER). In Table 2, identity mapping means $\omega = 1$, which represents some GAN-based work, ignoring the internal connection of frontal-profile face pairs, and only relying on the generator and the discriminator to produce results. Linear mapping: $\omega = 2\theta / \pi$ is a natural attempt and meets the design constraints. Besides,

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Verification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backbone</td>
<td>92.96</td>
</tr>
<tr>
<td>LARNet</td>
<td>98.84</td>
</tr>
<tr>
<td>LARNet+</td>
<td><strong>99.21</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gating Control Function</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity mapping: $\omega \equiv 1$</td>
<td>15.35</td>
</tr>
<tr>
<td>Linear mapping: $\omega = \theta / \pi$</td>
<td>9.68</td>
</tr>
<tr>
<td>Nolinear mapping: $\omega = \text{sigmoid}(4\theta / \pi - 1)$</td>
<td>8.45</td>
</tr>
<tr>
<td>PReLU</td>
<td>9.72</td>
</tr>
<tr>
<td>cReLU with OW</td>
<td>7.92</td>
</tr>
<tr>
<td>LARNet: $\omega =</td>
<td>\sin \theta</td>
</tr>
</tbody>
</table>
Table 3. Quantitative evaluation on the IJB-A dataset, where o.s. denotes the optimal setting, and f. is fine-tuning/refinement. Symbol ‘-’ indicates that the metric is not available for that protocol.

<table>
<thead>
<tr>
<th>Method</th>
<th>TAR@FAR=0.01</th>
<th>TAR@FAR=0.001</th>
<th>Rank-1 Acc.</th>
<th>Rank-5 Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al. (2016)</td>
<td>0.729</td>
<td>0.510</td>
<td>0.822</td>
<td>0.931</td>
</tr>
<tr>
<td>Pooling Faces (Hassner et al., 2016)</td>
<td>0.819</td>
<td>0.631</td>
<td>0.846</td>
<td>0.933</td>
</tr>
<tr>
<td>Multi Pose-Aware (AbdAlmageed et al., 2016)</td>
<td>0.787</td>
<td>—</td>
<td>0.846</td>
<td>0.927</td>
</tr>
<tr>
<td>DCNN Fusion (f.) (Chen et al., 2016)</td>
<td>0.838</td>
<td>—</td>
<td>0.903</td>
<td>0.965</td>
</tr>
<tr>
<td>PAMs (Masi et al., 2016a)</td>
<td>0.826</td>
<td>0.652</td>
<td>0.840</td>
<td>0.925</td>
</tr>
<tr>
<td>Augmentation+Rendered (Masi et al., 2016b)</td>
<td>0.886</td>
<td>0.725</td>
<td>0.906</td>
<td>0.962</td>
</tr>
<tr>
<td>Multi-task learning (Yin &amp; Liu, 2017)</td>
<td>0.787</td>
<td>—</td>
<td>0.858</td>
<td>0.938</td>
</tr>
<tr>
<td>TPE(f.) (Sankaranarayanan et al., 2017)</td>
<td>0.900</td>
<td>0.813</td>
<td>0.932</td>
<td>—</td>
</tr>
<tr>
<td>DR-GAN (Tran et al., 2017)</td>
<td>0.831</td>
<td>0.699</td>
<td>0.901</td>
<td>0.953</td>
</tr>
<tr>
<td>FF-GAN (Yin et al., 2017)</td>
<td>0.852</td>
<td>0.663</td>
<td>0.902</td>
<td>0.954</td>
</tr>
<tr>
<td>NAM (Yang et al., 2017)</td>
<td>0.921</td>
<td>0.861</td>
<td>0.938</td>
<td>0.960</td>
</tr>
<tr>
<td>Multicolumn (Xie &amp; Zisserma, 2018)</td>
<td>0.920</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>VGGFace2 (Cao et al., 2018b)</td>
<td>0.904</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Template Adaptation(f.) (Crosswhite et al., 2018)</td>
<td>0.939</td>
<td>—</td>
<td>0.928</td>
<td>—</td>
</tr>
<tr>
<td>DREAM (Cao et al., 2018a)</td>
<td>0.872</td>
<td>0.712</td>
<td>0.915</td>
<td>0.962</td>
</tr>
<tr>
<td>DREAM(E2E+retrain,f.) (Cao et al., 2018a)</td>
<td>0.934</td>
<td>0.836</td>
<td>0.939</td>
<td>0.960</td>
</tr>
<tr>
<td>FTL with 60K parameters (o.s.) (Yin et al., 2019)</td>
<td>0.864</td>
<td>0.744</td>
<td>0.893</td>
<td>0.947</td>
</tr>
<tr>
<td>PFEs (Shi &amp; Jain, 2019)</td>
<td>0.944</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>DebFace (Gong et al., 2020)</td>
<td>0.902</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Rotate-and-Render (Zhou et al., 2020)</td>
<td>0.920</td>
<td>0.825</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>HPDA (Wang et al., 2020)</td>
<td>0.876</td>
<td>0.803</td>
<td>0.84</td>
<td>0.88</td>
</tr>
<tr>
<td>CDA (Wang &amp; Deng, 2020)</td>
<td>0.911</td>
<td>0.823</td>
<td>0.936</td>
<td>0.957</td>
</tr>
<tr>
<td>LARNet</td>
<td>0.941</td>
<td>0.842</td>
<td>0.936</td>
<td>0.968</td>
</tr>
<tr>
<td>LARNet+</td>
<td>0.951</td>
<td>0.874</td>
<td>0.949</td>
<td>0.971</td>
</tr>
</tbody>
</table>

Our gating control function essentially acts as a filtering activation function, and we compare it with two widely used activation functions PReLU (He et al., 2015) and cReLU with OW (Balduzzi et al., 2017). Nonlinear mapping: $\omega = \text{sigmoid}(\frac{40}{\pi} - 1)$ is reported by DREAM (Cao et al., 2018a). Our method $\omega = |\sin \theta|$ still achieves the best result: EER 6.26%. This observation ascertains our design of exerting a higher degree of correction to a profile face, and has a better understanding about the effectiveness of our proposed LARNet for profile face recognition.

4.4. Quantitative Evaluation Results

We compare our method with more than 30 competitors published in the recent five years with respect to different metrics, which include template based, GAN, residual learning, and three-dimensional reconstruction, aiming at handling various tasks such as face search, face recognition, face verification, large pose recognition, etc. All numerical statistics are the best results obtained from original quota- tion, cross-reference, and experimental reproduction.

4.4.1. IJBA DATASET: VERIFICATION AND IDENTIFICATION WITH STATE OF THE ARTS

In this experiment, we evaluate our method on the challenging benchmark IJBA that covers full pose variations and complies with the original standard protocol. The evaluation metrics include popular True Acceptance Rate (TAR) at False Acceptance Rate (FAR) of 0.01 and 0.001 on the verification task, and Rank-1/Rank-5 recognition accuracy on the identification task. All methods employ the same MS1MV2 dataset and ResNet-50 backbone for training.

In Table 3, we compare with various state-of-the-art techniques. Our LARNet reaches 0.941 (TAR@FAR=0.01), and after refinement with end-to-end retraining, LARNet+ achieves a better performance with 0.951. They have surpassed the other existing methods by a large margin. Our method also brings significant improvement in a more challenging metric TAR@FAR=0.001, with the results of 0.842 and 0.874, respectively. Furthermore, for face identification, LARNet has an advantage in both Rank-1 Acc. (0.936) and Rank-5 Acc. (0.968), and LARNet+ pushes the result to a higher level: Rank-1 Acc. (0.949) and Rank-5 Acc. (0.971). Our method has achieved superior performance on both recognition and verification tasks.

4.4.2. CFP-FP DATASET: PROFILE FACE VERIFICATION CHALLENGE

We employ CFP-FP as the frontal profile face verification dataset with the protocol that the whole dataset is divided into 10 folds each containing 350 same and 350 not-same pairs of 50 individuals. All methods employ the same MS1MV2 dataset and ResNet-50 backbone for training. From Table 4, we observe that the face verification results of state-of-the-art face recognition models are basically at 94%+. It is worth noting that in 2020, a latest work named universal representation learning face (URFace) (Shi et al.,...
Table 4. Quantitative evaluation on the CFP-FP dataset, where o.s. denotes the optimal setting, and f. is fine-tuning/refinement.

<table>
<thead>
<tr>
<th>Method</th>
<th>Verification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SphereFace (o.s.+f.)</td>
<td>94.17</td>
</tr>
<tr>
<td>CosFace (o.s.)</td>
<td>94.40</td>
</tr>
<tr>
<td>ArcFace (o.s.+f.)</td>
<td>94.04</td>
</tr>
<tr>
<td>URFace (all modules, MS1MV2, o.s.)</td>
<td>98.64</td>
</tr>
<tr>
<td>Human-level</td>
<td>98.92</td>
</tr>
<tr>
<td>LARNet</td>
<td>98.84</td>
</tr>
<tr>
<td>LARNet+</td>
<td><strong>99.21</strong></td>
</tr>
</tbody>
</table>

Table 5. Quantitative evaluation on the general face recognition datasets: LFW, YTF, and CPLFW, where o.s. denotes the optimal setting, and f. is fine-tuning/refinement. Symbol '-' indicates that the metric is not available for that protocol. For fairness, the training network and dataset: ResNet-50 + CASIA-WebFace. Symbol '*' indicates the methods under different training designs. The evaluation metric is verification rate (%).

<table>
<thead>
<tr>
<th>Method</th>
<th>LFW</th>
<th>YTF</th>
<th>CPLFW</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUMAN-Individual</td>
<td>97.27</td>
<td>78.21</td>
<td></td>
</tr>
<tr>
<td>HUMAN-Fusion</td>
<td><strong>99.85</strong></td>
<td>-</td>
<td>85.24</td>
</tr>
<tr>
<td>DeepID (Sun et al., 2014)</td>
<td>99.47</td>
<td>93.20</td>
<td>-</td>
</tr>
<tr>
<td>Deep Face (Taigman et al., 2014)</td>
<td>97.35</td>
<td>91.4</td>
<td>-</td>
</tr>
<tr>
<td>VGG Face (Parkhi et al., 2015)</td>
<td>98.95</td>
<td>97.30</td>
<td>90.87</td>
</tr>
<tr>
<td>FaceNet (Schröff et al., 2015)</td>
<td>99.63</td>
<td>95.10</td>
<td>-</td>
</tr>
<tr>
<td>Baidu (Liu et al., 2015a)</td>
<td>99.13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Center Loss (Wen et al., 2016)</td>
<td>99.28</td>
<td>94.9</td>
<td>85.48</td>
</tr>
<tr>
<td>Range Loss (Zhang et al., 2017)</td>
<td>99.52</td>
<td>93.70</td>
<td>-</td>
</tr>
<tr>
<td>Marginal Loss (Deng et al., 2017)</td>
<td>99.48</td>
<td>95.98</td>
<td>-</td>
</tr>
<tr>
<td>SphereFace(o.s.) (Liu et al., 2017)</td>
<td>99.42</td>
<td>95.0</td>
<td>81.4</td>
</tr>
<tr>
<td>SphereFace+(o.s.)</td>
<td>99.47</td>
<td>-</td>
<td>90.30</td>
</tr>
<tr>
<td>CosFace(o.s.) (Wang et al., 2018a)</td>
<td>99.51</td>
<td>96.1</td>
<td>-</td>
</tr>
<tr>
<td>CosFace*(MS1MV2,R64, o.s.)</td>
<td>99.73</td>
<td>97.6</td>
<td>-</td>
</tr>
<tr>
<td>Arcface(o.s.) (Deng et al., 2019)</td>
<td>99.53</td>
<td>-</td>
<td>92.08</td>
</tr>
<tr>
<td>ArcFace*(MS1MV2,R100,f.)</td>
<td>99.83</td>
<td><strong>98.02</strong></td>
<td>95.45</td>
</tr>
<tr>
<td>Ours: LARNet</td>
<td>99.36</td>
<td>96.55</td>
<td>95.51</td>
</tr>
<tr>
<td>Ours: LARNet+</td>
<td>99.71</td>
<td>97.63</td>
<td><strong>96.23</strong></td>
</tr>
</tbody>
</table>

(2020) has reached an astonishing 98.64% under the training of the MS1MV2 dataset and the auxiliary learning of a large number of modules, such as variation augmentation, confidence-aware identification loss, and multiple embeddings. However, the result of our LARNet has reached 98.84%, which outperforms almost all competitors. Regarding further advanced LARNNet+, to our best knowledge, the 99.21% performance is the first to surpass the reported human-level performance (98.92%) on the CFP-FP dataset.

4.4.3. LFW, YTF, and CPLFW Datasets: General Face Recognition

To get a better understanding about our LARNet, we conduct a more in-depth comparison on general face recognition. LFW and YTF datasets are the most widely used benchmark for unconstrained face verification on images and videos. In this experiment, we follow the unrestricted with labelled outside data protocol to report the performance. CPLFW emphasizes pose difference to further enlarge intra-class variance. All methods employ the same CASIA-WebFace dataset and ResNet-50 backbone, while the results of some different experimental designs (marked with ‘*’*) are also reported.

From Table 5, we find that because the LFW dataset is too small, almost all methods can achieve 99%+. Although the meaning is very weak, the result 99.74% of our method is also at the forefront, only lower than human-level 99.85% and Arcface* 99.83%. It is worth noting that Arcface* is trained with the MS1MV2 dataset (5.8M) and ResNet-100, while Arcface with CASIA-WebFace (100K) and shallower ResNet-50 (the same as ours) only reaches 99.53% and is inferior to ours. We provide this more comparative result for an extensive study. For the video sampled dataset YTF, there exists the same situation, and our LARNet and LARNet+ are still superior to all ResNet-50 based face recognition methods and slightly inferior to deeper ResNet-100 based Arcface*. We also introduce a more challenging CPLFW dataset with large poses, which has more realistic consideration on pose intra-class variations. Our method achieves the best results 95.51% and 96.23%, respectively.

5. Conclusions

We proposed the Lie Algebra Residual Network (LARNet) to boost face recognition performance. First, we presented a novel method with Lie algebra theory to explore how face rotation in the 3D space affects the deep feature generation process, and proved that a face rotation in the image space is equivalent to an additive residual component in the deep feature space. Furthermore, we designed a gating control function, which is derived on the foundation of Lie algebra to learn rotation magnitude and control the impact of the residual component on the feature learning process. Moreover, we provided the results of ablation studies to validate the effectiveness of our Lie algebraic deep feature learning. The comprehensive experimental evaluations demonstrate the superior performance of the proposed LARNet over the state-of-the-art methods on frontal-profile face verification, face identification, and general face recognition tasks. In future work, we will continue to pay attention to the interpretability of Euclidean transformation in other CNNs, and intend to explore more mathematical tools to decouple potential geometric properties hidden in the image space.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Nos. 12022117, 61872354 and 61772523), the Beijing Natural Science Foundation (No. Z190004), and the Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202023).
References


LARNet: Lie Algebra Residual Network for Face Recognition


Masi, I., Tran, A. T., Hassner, T., Leksut, J. T., and Medioni, G. Do we really need to collect millions of faces for effective face recognition? In European Conference on Computer Vision (ECCV), pp. 579–596, 2016b.


